Sparse Representation Classification Beyond $\ell 1$ Minimization and the Subspace Assumption

Cencheng Shen[a], Li Chen[a], Carey E. Priebe[a,*]

[a]*Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218*

Abstract

The sparse representation classifier (SRC) proposed in [1] has recently gained much attention from the machine learning community. It makes use of $\ell 1$ minimization, and is known to work well for data satisfying a subspace assumption. In this paper, we use the notion of class dominance as well as a principal angle condition to investigate and validate the classification performance of SRC, without relying on $\ell 1$ minimization and the subspace assumption. We prove that SRC can still work well using faster subset regression methods such as orthogonal matching pursuit and marginal regression, and its applicability is not limited to data satisfying the subspace assumption. We illustrate our theorems via various real data sets including face images, text features, and network data.

Keywords: sparse representation classification, $\ell 1$ minimization, orthogonal matching pursuit, marginal regression, class dominance, principal angle

1. Introduction

Recently there is a surge in utilizing the sparse representation and regularized regression for many machine learning tasks in computer vision and pattern recognition. Applications include [2], [1], [3], [4], [5], [6], among many others. In this paper, we concentrate on one specific but profound application – the sparse representation classification (SRC), which is proposed by [1] and exhibits state-of-the-art performance for robust face recognition.

For the classification task, denote $x \in \mathcal{R}^m$ as the testing observation and $\mathcal{X} \in \mathcal{R}^{m \times n}$ as the matrix of training data with all columns pre-scaled to unit-norm. Each column of $\mathcal{X}$ is denoted

*Corresponding author

Email addresses: cshen6@jhu.edu (Cencheng Shen), lchen87@jhu.edu (Li Chen), cep@jhu.edu (Carey E. Priebe)

as x_i for $i = 1, \ldots, n$, representing a training observation with a known class label $y_i \in [1, \ldots, K]$. The sparse representation classifier consists of two steps: for each testing observation x, first it finds a sparse representation β such that $x = \mathcal{X}\beta + \epsilon$; then the class of the testing observation is determined by $g(x) = \arg\min_{k=1,\ldots,K} \|x - \mathcal{X}\beta_k\|_2$, where $g(\cdot) : \mathcal{R}^m \to \{1, \ldots, K\}$ is the classifier, and β_k takes the values from β that are associated with data of class k, i.e., $\beta_k(i) = \beta(i)$ if $y_i = k$, 0 otherwise. Denoting the true but unknown class of x as $y \in [1, \ldots, K]$, SRC correctly finds the true label if $g(x) = y$. This classifier has been numerically shown to work well and be robust against occlusion and contamination on face images, and argued to be better than nearest-neighbor and nearest-subspace rules in [1].

Clearly finding an appropriate sparse representation is the crucial step of SRC, which is intrinsically subset regression, i.e., apply certain method to select a subset of data $\mathcal{X}_s \in \mathcal{X}$, and then take the corresponding regression vector as the sparse representation β. Most works on sparse representation have been using regularized regression methods to achieve the sparsity, for which $\ell 1$ minimization/Lasso are very popular choices due to their theoretical justifications [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], etc. The literature in $\ell 1$ minimization and Lasso are more than abundant, and usually emphasizes how the $\ell 1$ regularization can help recover the most sparse model. But how the regularization may help the subsequent inference is usually a difficult question to answer in practice, and the role of $\ell 1$ minimization is not entirely clear for this particular classification task.

The initial motivation for [1] to use $\ell 1$ minimization is its equivalence to $\ell 0$ minimization (i.e., sparse model recovery) under various conditions, such as the incoherence condition [9] or restricted isometry property [11]. Namely if the testing observation x does have a unique and correct most sparse representation β (correct in the sense of $g(x) = y$) with respect to the training data, then assuming proper conditions are satisfied, $\ell 1$ minimization is an ideal choice in the SRC framework. But the sample training data are usually correlated, which violates many theoretical conditions including incoherence and restricted isometry; furthermore, the classification task requires only the recovered β to be mostly associated with data of the correct class rather than one uniquely correct solution, so there usually exists infinite correct β which are not the most sparse solution.

Towards this direction, it is argued in [1] that if data of the same class lie in the same subspace while data of different classes lie in different subspaces (called the subspace assumption henceforth), then most data selected by $\ell 1$ minimization should be from the correct class, thus yielding good

classification performance in SRC. Since face image data under varying lighting and expression roughly satisfy the subspace assumption [18], [19], they further argue that SRC is applicable to face images. Indeed, based on the subspace assumption, [6] derives a theoretical condition for $\ell 1$ minimization to do perfect variable selection, i.e., all selected training data are from the correct class. This indicates that sparse representation is a valuable tool with $\ell 1$ minimization under the subspace assumption. But the subspace assumption assumes a low-dimensional structure for data of the same class, which does not always hold and is difficult to validate in practice.

Despite many applications of and investigations into sparse representation, the intrinsic properties and mechanisms of SRC are still not well understood, and there exist evidence [20], [21], [22], [23] that neither $\ell 1$ minimization nor the subspace assumption are necessary in the SRC framework. In particular, [21] and [23] argue that it is actually the classification step (namely $g(x) = \arg\min_{k=1,\ldots,K} \|x - \mathcal{X}\beta_k\|_2$) that is most effective in SRC; they call it the collaborative representation, and support their claims through many numerical examples. Our previous work applies SRC to vertex classification [24], which also achieves good performance for network data without $\ell 1$ minimization or the subspace assumption.

To deepen our understanding, in this paper we target two important questions related to SRC. First, is the subspace assumption a necessity for SRC to perform well? And if not, when and how is SRC applicable with theoretical performance guarantees? Second, despite the popularity of $\ell 1$ minimization, is this the optimal approach to do variable selection for SRC? Can we use other faster subset regression methods such as orthogonal matching pursuit (OMP) [25], [26] and marginal regression [27], [28]?

With these two target questions in mind, this paper is organized as follows. In Section 2 we review the SRC framework and three subset regression methods including $\ell 1$ homotopy, OMP, and marginal regression. Section 3 is the main section. In subsection 3.1 we first relate SRC to a notion we call class dominance on the sample data. Then, based on class dominance, in subsection 3.2 we state a principal angle condition on the data distribution that is sufficient for the classification consistency of SRC. In particular, our theorems largely explain the success of SRC, are still valid when $\ell 1$ minimization is replaced by OMP or marginal regression, and can help identify data models for SRC to work well without requiring the subspace assumption. Our results make SRC more appealing in terms of theoretical foundation, computational complexity and general applicability, and are illustrated via various real data sets including face images, text features, and network data

in Section 4. We conclude the paper in Section 5, with all proofs relegated to Section 6.

2. Sparse Representation Review

2.1. The SRC Algorithm

We first summarize the SRC algorithm using $\ell 1$ minimization in Algorithm 1, which consists of the subset regression step and the classification step.

Algorithm 1 Sparse representation classification by $\ell 1$ minimization

Input: An $m \times n$ matrix $\mathcal{X}$, where each column x_i represents a training observation with a known label $y_i \in [1, \ldots, K]$. An $m \times 1$ testing observation x with its true label y being unknown. Unless mentioned otherwise, we always assume each column of x and $\mathcal{X}$ are pre-scaled to unit norm, and $\mathcal{X}$ is not orthogonal to x (otherwise β is always the zero vector).

1. **Find a sparse representation of x by $\ell 1$ minimization**:

$$\text{Solve: } \beta = \arg\min \|\beta\|_1 \text{ subject to } \|x - \mathcal{X}\beta\|_2 \leq \epsilon. \tag{1}$$

2. **Classify x by the sparse representation β**:

$$g(x) = \arg \min_{k=1,\ldots,K} \|x - \mathcal{X}\beta_k\|_2, \tag{2}$$

break ties deterministically. For each entry of β_k, $\beta_k(i) = \beta(i)$ if $y_i = k$, 0 otherwise.

Output: The assigned label $g(x)$.

Solving Equation 1 by $\ell 1$ minimization is the only computational costly part of SRC. There are many possible methods to solve $\ell 1$ minimization, see in [29], [30], [5], [6], among which we use the $\ell 1$ homotopy method for subsequent analysis and numerical experiments. This method is based on a polygonal solution path [31], [32] and can also be used for Lasso and least angle regression [7], [10].

Alternatively, OMP is a greedy approximation of $\ell 1$ minimization and is equivalent to forward stepwise regression; it gains its popularity in sparse recovery due to its reduced running time and certain theoretical guarantees [26], [33], [34], [35]. Furthermore, OMP is quite similar to $\ell 1$ homotopy in the implementation, and there exist many extensions of OMP [36], [37], [38].

As for marginal regression, it is probably the simplest and fastest way to do subset regression, and it has been studied and applied successfully in many areas. Despite its simplicity, it has been shown to work well for variable selection in high-dimensional data comparing to Lasso [27], [28],

[39], is particularly popular in ultra-high-dimensional screening [40], [41], [42], and has been applied to sparse representation as well [43]. We can always use OMP or marginal regression to find the sparse representation β in step 1, rather than solving Equation 1 by $\ell 1$ minimization. In the next subsection we compare $\ell 1$ homotopy, OMP, and marginal regression in more detail.

Note that the constraint in Equation 1 can be replaced by $x = \mathcal{X}\beta$ in a noiseless setting, but usually ϵ is required in order to achieve a more parsimonious model when dealing with high-dimensional or noisy data. This model selection problem, i.e., the choice of ϵ or more generally the sparsity level of subset regression, is a difficult problem intrinsic to most subset regression methods. We will explain this issue from the algorithmic point of view in the next subsection.

2.2. $\ell 1$ Homotopy, OMP, and Marginal Regression

As $\ell 1$ homotopy can be treated as an extension of OMP, and marginal regression is very simple, we only list the OMP algorithm in detail in Algorithm 2.

Algorithm 2 Use orthogonal matching pursuit to solve Step 1 of SRC

Input: The training data $\mathcal{X}$, the testing observation x, and a specified iteration limit s and/or a residual limit ϵ.

Initialization: The residual $r_0 = x$, iteration count $t = 1$, and the selected data $\mathcal{X}_0 = \emptyset$.

1. Find the index i_t such that $i_t = \arg\max_{i=1,\ldots,n} |x_i' r_{t-1}|$, where x_i is the ith column of $\mathcal{X}$ and $'$ is the transpose sign. Break ties deterministically, and add x_{i_t} into the selected data so that $\mathcal{X}_t = [\mathcal{X}_{t-1} \;\; x_{i_t}]$.

2. Update the regression vector β with respect to $\mathcal{X}_t$, i.e., calculate the orthogonal projection matrix $P_t = \mathcal{X}_t \mathcal{X}_t^{-}$ with $\mathcal{X}_t^{-}$ being the pseudo-inverse, and let $\beta = P_t x$. Then update the regression residual as $r_t = (I - P_t)x$.

3. If $t = s$, or $|r_t| < \epsilon$, or $|\mathcal{X}' r_t| \leq \epsilon 1_{n \times 1}$ entry-wise, stop and let $s = t$; else increment t.

Output: $\mathcal{X}_s$ and β. Note that the sparse representation β can be enlarged from an $s \times 1$ vector to an $n \times 1$ vector based on the relative positions of $\mathcal{X}_s$ in $\mathcal{X}$.

The idea of OMP is the same as forward selection: at each iteration OMP finds the column that is most correlated with the residual, and then re-calculates the regression vector by projecting x onto the selected sub-matrix $\mathcal{X}_t$. When the iteration limit is reached, or the residual is small enough, or the residual is almost orthogonal to the training data, OMP stops.

The $\ell 1$ homotopy method is the same as OMP in terms of the data selection, but it has an extra data deletion step and a different updating scheme. Conceptually, the homotopy path seeks $\beta = \min_\beta \|x - \mathcal{X}\beta\|_2/2 + \lambda\|\beta\|_1$ iteratively by reducing λ from a positive number to 0, which is proved to solve the $\ell 1$ minimization problem and can also be used for the Lasso regression. More details can be found in [10], [30]. Our experiments use the homotopy algorithm implemented by S. Asif and J. Romberg [1].

The marginal regression method does not involve any iteration; it simply chooses s columns out of $\mathcal{X}$ that are most correlated with the testing observation x, and calculates β to be the regression vector with respect to the selected $\mathcal{X}_s$. Because marginal regression is a non-iterative method, it enjoys a superior running time complexity comparing to others: for the data selection step, it takes only $O((m + \min(s, \log n))n)$ while OMP needs $O(mns)$; and for small s marginal regression is much faster than full regression (i.e., the usual $\ell 2$ minimization using full training data).

Clearly the three subset regression methods may yield different $\mathcal{X}_s$ and thus different β, but they always coincide at $s = 1$, which is an important fact for the later proof. Another useful observation is that $\mathcal{X}_s$ is always full rank when using $\ell 1$ homotopy or OMP (otherwise they stop), but this is not necessarily the case when using marginal regression after certain s. In the main section we will show that under a principal angle condition on the data model, all three methods can have the same asymptotic inferential effect, even though their sparse representation β may be different.

Note that the model selection problem is inherent in the stopping criteria, and the stopping criteria used in Algorithm 2 are commonly applied in subset regression. For example, [33] only specifies the iteration limit s to stop OMP, which is suitable when the testing observation is perfectly recoverable; [1] stops $\ell 1$ minimization for small residual $|r_t| < \epsilon$, which is more practical for real data, but a good choice of ϵ may be data-dependent; the almost orthogonal criterion (i.e., $|\mathcal{X}' r_t| \leq \epsilon 1_{n\times 1}$) has been used in [34], [35] for OMP to work well for sparse recovery; and other stopping criteria are also possible, such as Mallows's C_p. As model selection does not affect the main theoretical results, we do not delve into this topic; but its finite-sample inference effect for real data is often difficult to quantify, so in the numerical experiments we always plot the SRC error with respect to various sparsity levels while setting ϵ to be effectively zero, in order to give a fair evaluation of SRC for all possible models up to a certain limit.

[1] `http://users.ece.gatech.edu/~sasif/homotopy/`

3. Main Results

Let us introduce some notations before proceeding: $\mathcal{X}$ denotes the training data matrix of size $m \times n$, $\mathcal{X}_s$ denotes the selected sub-matrix of size $m \times s$ by subset regression, $\mathcal{X}_k$ denotes the sub-matrix of $\mathcal{X}_s$ whose columns are associated with class k, $\mathcal{X}_{-k}$ denotes the sub-matrix of $\mathcal{X}_s$ whose columns are not of class k. Furthermore, β represents the regression vector or sparse representation with respect to $\mathcal{X}_s$ or $\mathcal{X}$, which may be an $s \times 1$ vector or $n \times 1$ vector depending on the context, i.e., we use $\mathcal{X}_s\beta$ and $\mathcal{X}\beta$ interchangeably, where the former is the $s \times 1$ regression vector and the latter is the $n \times 1$ sparse representation; they only differ in zero entries. β_k equals β except every entry not associated with class k is 0, and $\beta_{-k} = \beta - \beta_k$; and similar to β, their size may be different depending on the context by shrinking or expanding the zero entries.

3.1. Class Dominance in the Regression Vector

We first define class dominance and positive class dominance for given regression vector and given sample data. They are not only important catalysts between the principal angle condition and the theoretical SRC optimality, but also crucial components underlying the empirical success of SRC as shown in the numerical section.

Definition. *Given β and the testing observation x and the training data $\mathcal{X}$, we say class y dominates β if and only if $\|\mathcal{X}\beta_{-y}\|_2 < \|\mathcal{X}\beta_y\|_2$.*

We say that class y positively dominates the regression vector β if and only if $\|\mathcal{X}\beta_y\|_2 \leq \|\mathcal{X}\beta_{-k}\|_2$ for all $k \neq y$.

Note that we say (positive) class dominance holds if and only if the correct class (positively) dominates the sparse representation of the testing observation.

For any given β, class dominance and positive class dominance together are sufficient for correct classification of SRC, formulated as follows.

Theorem 1. *Given β, (x, y) and $\mathcal{X}$, class y dominance implies $g(x) = y$ for SRC if class y also positively dominates β.*

If positive class dominance does not hold, class dominance itself is not sufficient for $g(x) = y$.

Although class dominance cannot guarantee correct classification in SRC, it is closely related to positive class dominance and can lead to the latter in many scenarios. The next corollary is an example.

Corollary 1. *Suppose $K = 2$, or the data are non-negative and the regression vector β is constrained to be non-negative.*

Then given β and the sample data, class dominance implies positive class dominance, in which case class dominance alone is sufficient for $g(x) = y$ in SRC.

Despite the limitations of Corollary 1, two-class classification problems are common; real data are often non-negative; and the non-negativity constraint is very useful in subset regression, such as the non-negative OMP [44] and the non-negative least squares [45], [46]. In fact, the condition in Corollary 1 can be further relaxed. For example, if the dominance magnitude is large enough (i.e., $c\|\mathcal{X}\beta_{-y}\|_2 < \|\mathcal{X}\beta_y\|_2$ for some $c \geq 1$) and the negative entries of β are properly bounded, then class dominance still implies positive class dominance and is sufficient for $g(x) = y$ in SRC.

This indicates that class dominance is usually sufficient for correct classification, unless the negative entries of β are too large. Indeed in the numerical section we observe that class dominance nearly always implies $g(x) = y$, even though the non-negative constraint is not used in subset regression; and the class dominance error is usually close to the classification error. In the next subsection, we make use of class dominance to identify a principal angle condition on the data model, so that SRC can be a consistent classifier without requiring $\ell 1$ minimization and the subspace assumption.

Note that the concept of class dominance appears similar to block sparsity and block coherence [47], [48], [49]. But block sparsity and block coherence are used to guarantee that the fewest number of blocks/classes of data are used in the sparse representation, which is not directly related to correct classification; while our class dominance is defined for the correct class of data to dominate the sparse representation, which can lead to correct classification.

3.2. Classification Consistency of SRC

In this subsection we formalize the probabilistic setting of classification based on [50]. Suppose $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} F_{XY}$, where $(X, Y) \in \mathcal{R}^m \times \{1, \ldots, K\}$ is the random variable pair generating the testing observation and its class (x, y), (X_i, Y_i) are the random variables generating the training pair (x_i, y_i) for $i = 1, \ldots, n$. Note that the prior probability of the data being in each class k should be nonzero.

The SRC error is denoted as $L = Prob(g(X) \neq Y | (X_1, Y_1), \ldots, (X_n, Y_n))$ for the SRC classifier $g : \mathcal{R}^m \to \{1, \ldots, K\}$. We always have $L \geq L^*$, where L^* is the optimal Bayes error. For SRC to

achieve consistent classification, it is equivalent to identify a sufficient condition on F_{XY} so that $L \to L^*$ as $n \to \infty$. We henceforth consider the case that $L^* = 0$ so that $L \to 0$ implies SRC is asymptotically optimal.

Based on this probabilistic setting and the previous subsection on class dominance, the SRC error can be decomposed by conditioning on class dominance:

$$\begin{aligned} L &= Prob(\text{class dominance}) \times Prob(g(X) \neq Y | \text{class dominance}) \\ &\quad + Prob(\text{class dominance fails}) \times Prob(g(X) \neq Y | \text{class dominance fails}). \\ &= P_D \times P_1 + (1 - P_D) \times P_2, \end{aligned} \tag{3}$$

where P_D denotes the class dominance probability, P_1 denotes the conditional classification error given class dominance, and P_2 denotes the conditional classification error when class dominance fails. Clearly the class dominance probability P_D depends on both F_{XY} and the subset regression method; moreover, Corollary 1 indicates that $P_1 = 0$ when X is non-negative and β is derived under the non-negative constraint, which approximately holds throughout our numerical experiments without the non-negative constraint.

So for SRC to perform well, it suffices to find a condition on F_{XY} so that P_D is close to 1, then the SRC error L can be close to 0; and for SRC to be optimal beyond $\ell 1$ minimization and the subspace assumption, the condition should be as simple and as general as possible, not requiring the subspace assumption, yet still achieving class dominance almost surely for most subset regression methods.

First we state an auxiliary condition to ensure class dominance for given $\mathcal{X}_s$ of full rank, which serves as a starting point for the later results.

Theorem 2. *Given β, (x, y) and any selected data matrix $\mathcal{X}_s$ of full rank, class dominance holds if and only if*

$$\theta(x, \mathcal{X}_y \beta_y) < \theta(x, \mathcal{X}_{-y} \beta_{-y}), \tag{4}$$

where $\theta(x, \cdot)$ denotes the principal angle between x and $\cdot$.

Therefore, when Equation 4 holds for the selected sub-matrix $\mathcal{X}_s$, class y dominates the sparse representation. We can convert this condition into the probabilistic setting as follows.

Theorem 3. *Under the probabilistic setting, we define the principal angle condition as follows: for fixed (x, y), there exists a constant $c_{xy} \in [0, \pi/2)$ such that $\theta(x, X_1) \leq c_{xy} | Y_1 = y$ almost surely and $\theta(x, [X_1, \cdots, X_s]) > c_{xy} | Y_i \neq y, i = 1, \ldots, s$ almost surely.*

Denote q as the probability that the principal angle condition holds for $(X, Y) \sim F_{XY}$. Then the class dominance probability P_D is asymptotically no less than q, for $\mathcal{X}_s$ derived by $\ell 1$ minimization at any given $s \geq 1$.

Namely, class dominance holds if the within-class data are close while the between-class data are far away in terms of the principal angle. By Equation 3 and Corollary 1, it is clear that the principal angle condition can lead to SRC optimality, which we state as a corollary.

Corollary 2. *Suppose both the principal angle condition in Theorem 3 and the condition in Corollary 1 hold with probability q for $(X, Y) \sim F_{XY}$. Then the SRC error using $\ell 1$ minimization satisfies $L \leq 1 - q$ asymptotically.*

Thus if $q \to 1$ (i.e., all possible (x, y) in the support of F_{XY} satisfy the principal angle condition), SRC is asymptotically optimal with $L \to 0$.

Thus this condition does not explicitly rely on the subspace assumption, yet still leads to optimal classification and can be used to validate SRC applicability on general data models. At $s = 1$, the principal angle condition can be easily validated by the nearest neighbor based on principal angle/correlation. But for large s, the between-class principal angle is more difficult to check: if the subspace assumption holds, the principal angle between one observation and s observations of other classes are usually bounded below, so the condition holds as long as the within-class angle is small; if the subspace assumption does not hold, the principal angle condition at large s may not hold even for well-separated data.

Therefore it is sometimes useful to prove the principal angle condition together with the non-negative constraint in Corollary 1. Because $\cos \theta(x, [X_1, \cdots, X_s])$ equals the correlation between x and a linear combination of $[X_1, \cdots, X_s]$, we can require the correlation between x and any non-negative linear combination of $[X_1, \cdots, X_s]$ to be small instead of $\theta(x, [X_1, \cdots, X_s])$ to be large; then Corollary 2 still holds. One such application is illustrated in [24] for the adjacency matrix.

The proof of Theorem 3 can be adapted to any of $\ell 1$ minimization, OMP, and marginal regression, which yields the next corollary.

Corollary 3. *When $\ell 1$ minimization is replaced by OMP in the SRC framework, Theorem 3 and Corollary 2 still hold.*

Furthermore, if we constrain the sparsity level s such that $\mathcal{X}_s$ selected by marginal regression is full rank (which is always possible up to certain s), or the original data $\mathcal{X}$ itself is full rank, then Theorem 3 and Corollary 2 also hold for SRC using marginal regression or full regression.

Therefore, not only can OMP and marginal regression be used in SRC, so can full regression. However, for real data it is quite common that the full training data matrix $\mathcal{X}$ is either rank deficient or very close to rank deficient (i.e., having singular values very close to 0).

So far our principal angle condition in Theorem 3 is quite restrictive, especially $\theta(x, X_1) < c_y | Y_1 = y$ almost surely, as it requires data of the correct class to be always close. This can be relaxed as long as some data of the correct class are close enough to the testing observation, at the cost of treating far away data of the correct class as data of another class.

Corollary 4. *Under the probabilistic setting, suppose we extend the principal angle condition as follows: for fixed (x, y), there exists a constant $c_{xy} \in [0, \pi/2)$ such that*

$$X_1 | (Y_1 = y) = X_{+1} I_{\theta(x, X_1) \leq c_{xy}} + X_{-1} I_{\theta(x, X_1) > c_{xy}}, \tag{5}$$

where $I.$ is the indicator function, and $\theta(x, [X_1, \cdots, X_s]) > c_{xy} | \text{either } Y_i \neq y \text{ or } X_i \sim X_{-1}, i = 1, \ldots, s$ almost surely.

Then Theorem 3, Corollary 2 and Corollary 3 still hold. Note that the previous principal angle condition in Theorem 3 is now a special case with $Prob(I_{\theta(x, X_1) > c_{xy}}) = 0$.

Overall, our results in this subsection can be interpreted as demonstrating that for any given model, if the within-class principal angle can be small while the between-class principal angle is always large, then the correct class is likely to dominate the sparse representation, and SRC will succeed in the classification task. The principal angle condition here is similar to the condition in [6]: their condition is applied on given sample data while we focus more on the distribution; and their condition explicitly requires the subspace assumption and $\ell 1$ minimization while we do not.

Furthermore, we have addressed the two questions regarding SRC in the introduction: our principal angle condition can be used to check whether SRC is applicable to a given data model without the subspace assumption, for which class dominance plays a crucial role for correct classification;

the theorems also indicate that SRC should perform similarly for any of the aforementioned three subset regression methods. They are all reflected in the numerical section.

4. Numerical Experiments

In this section we apply the sparse representation classifier to various simulated and real data sets using $\ell 1$ homotopy, OMP, and marginal regression, and illustrate how our theoretical derivation of SRC is closely related to its numerical performance.

All experiments are carried out by hold-out validation, and for each data set we always randomly split the data in half for training and testing. Then we estimate the SRC error, the class dominance error, the SRC error given class dominance, and the SRC error when class dominance fails, i.e., the estimates of L, $1 - P_D$, P_1 and P_2 in Equation 3. To give a fair evaluation and account for possible early termination by various model selection criteria, the errors are always calculated against the sparsity level from $s = 1, \ldots, 100$, i.e., we re-calculate the regression vector and re-classify the testing observation for each s.

We also add k-nearest-neighbor (kNN) and linear discriminant analysis (LDA) for benchmark purposes of the classification error. They are calculated against the projection dimensions, i.e, we linearly project the data into dimension $d = 1, \ldots, 100$ by principal component analysis (or spectral embedding if the input is a dissimilarity/similarity matrix), and apply 9-nearest-neighbor (9 is just an arbitrary choice) and LDA on the projected data.

In all examples, the above procedure is repeated for 100 Monte Carlo replicates with the mean errors presented.

4.1. Face Images

We first apply SRC to two face image data sets, one of which is also used by [1] to show the empirical advantage of SRC.

The Extended Yale B database has 2414 face images of 38 individuals under various poses and lighting conditions [51], [52]. These images are further re-sized to 32×32 for our experiment. Half of the data is used for training and the other half for testing, so $m = 1024$, $n = 1207$, and $K = 38$. We show the mean errors after 100 Monte Carlo runs in Figure 1.

The CMU PIE database has 11554 images of 68 individuals under various poses, illuminations and expressions [53]. We also use the size 32×32 re-sized images for classification, so $m = 1024$, $n = 5777$, and $K = 68$. The mean errors are shown in Figure 2.

The top left panel of each figure shows the SRC error, and we observe that the error rates for different subset regression methods are very similar. The best error achieved for Extended Yale B database is 0.0207 by OMP, and the best error for CMU PIE is 0.0239 by $\ell 1$ minimization. Note that SRC by full regression achieves a mean error of 0.0606 and 0.0442 respectively, which is a bit worse than subset regression. As for kNN and LDA, their error rates are always greater than 0.1 in both data sets, which are not shown in order to better compare the SRC errors.

For both data sets, the top right panel shows the class dominance error $1 - P_D$, which is slightly higher than the SRC errors L; their difference is less than 0.05. The classification error given class dominance satisfies $P_1 < 0.01$ for all three subset regression methods and all sparsity levels, which is not shown by figure. The bottom panel shows P_2, which is much higher than L and P_1.

Those additional panels demonstrate that class dominance largely explains the success of SRC for face images, and the testing data can only be misclassified due to the failure of class dominance. Since all three subset regression methods achieve almost zero errors given class dominance, it is also the main reason that all methods have similar classification errors in the top left panel.

4.2. Wikipedia Data

Next we apply SRC to our Wikipedia documents with text and network features. We collect 1382 English documents from Wikipedia based on the 2-neighborhood of the English article "algebraic geometry", then form an adjacency matrix based on the documents' hyperlinks and a text feature distance matrix based on latent semantic analysis [54] and cosine distance. The data is available on our website [2]; and other examples of applying SRC to do vertex classification in graphs can be found in [24].

There are five classes in total for the documents, and both data sets are of size 1382×1382 (because the network data is an adjacency matrix, and the text feature data is a cosine distance matrix). Splitting half columns for training and the other half columns for testing, we have $m = 1382$, $n = 691$, and $K = 5$. The numerical performance is shown in Figure 3 and Figure 4 for the

[2] `http://www.cis.jhu.edu/~cshen/`

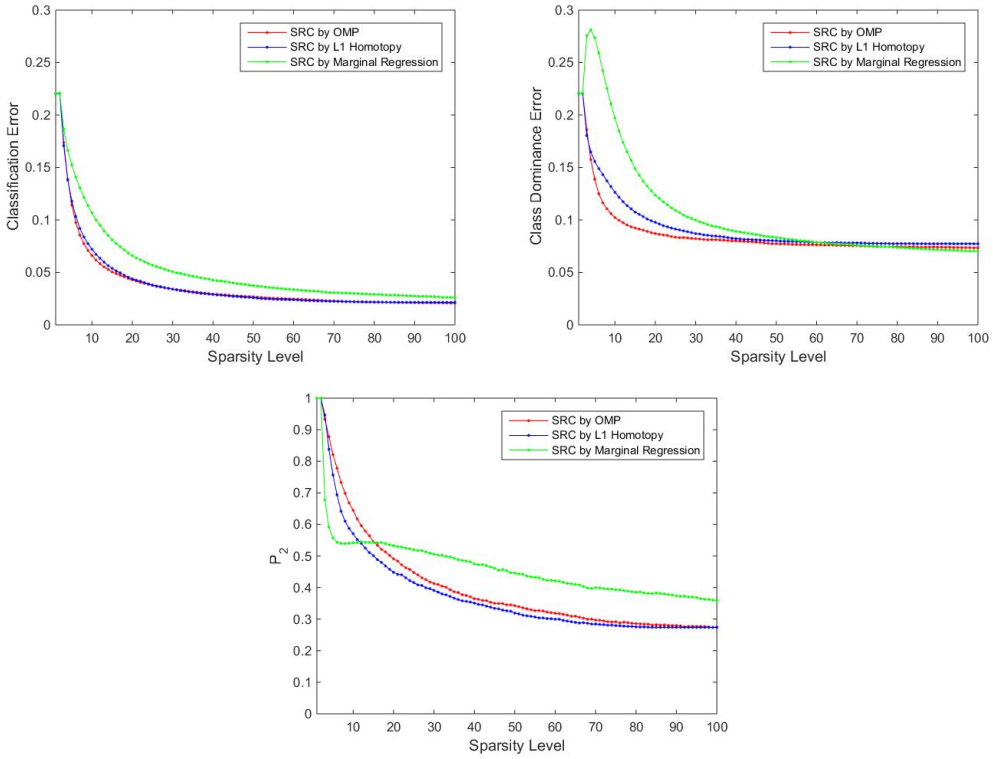

Figure 1: SRC for Extended Yale B Database

text and network data respectively. As the input data is a dissimilarity/similarity matrix, we use spectral embedding for projection prior to applying kNN and LDA.

The overall interpretation is similar to the face images: SRC performs quite well and is stable throughout different sparsity levels and different subset regression methods; the class dominance error $1 - P_D$ is higher than the SRC error L (for text data they are quite close; but for network data they are less close as sparsity level increases); $P_1 < 0.007$ in this example, indicating class dominance is crucial for correct classification; and P_2 is close to the chance line and much higher than L and P_1.

Note that the SRC classification errors for text features are lower than the network counterparts, because the text features should be more informative than the network data; we also observe that SRC becomes slightly inferior to LDA at large projection dimension d for text features, which is not the case for the adjacency matrix. This is probably because the cosine distance is a particularly

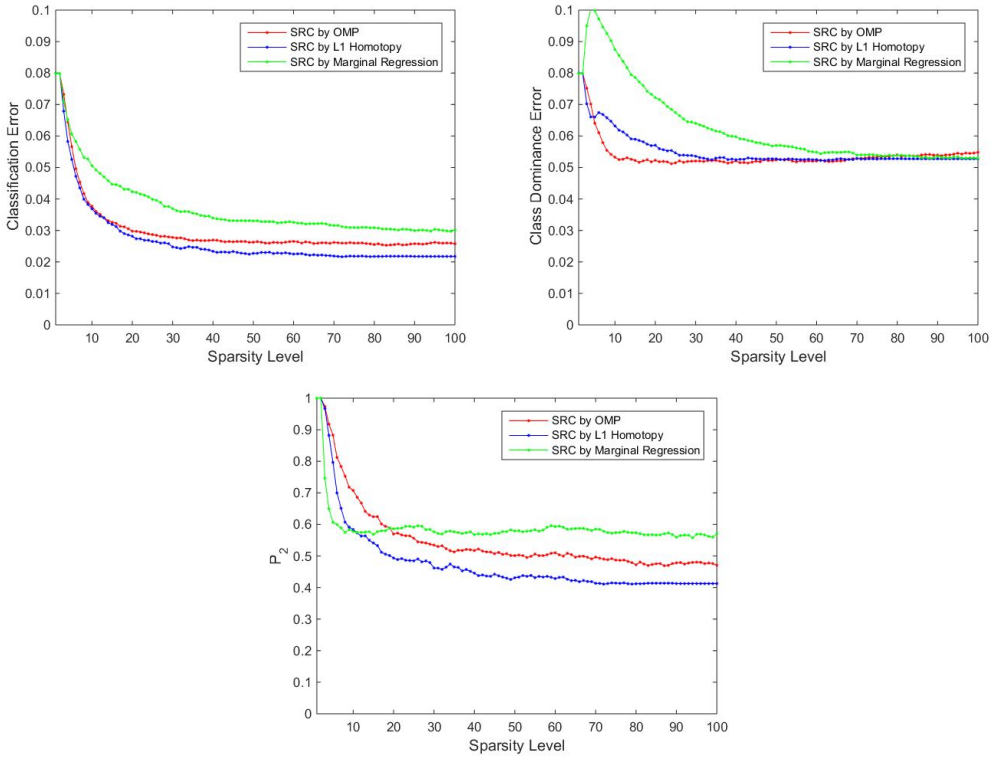

Figure 2: SRC for CMU PIE Database

suitable distance measure for text data [55], [56], thus allowing LDA to do better at proper projection dimensions. This phenomenon also holds for the face images in the previous subsection: even though LDA performs much worse than SRC on the raw data, LDA can achieve similar error rates as SRC for appropriate transformations of those images [57], [58].

5. Conclusion

In this paper we investigate the sparse representation classifier, which becomes very popular recently due to its empirical success for real data. In order to better understand the theory behind this method, we focus on the regression and classification steps of this method, and develop the notion of class dominance and principal angle condition. Our derivation establishes a theoretical foundation of sparse representation from a different point of view from current literature, as well as implying that $\ell 1$ minimization and the subspace assumption may not be crucial for SRC, which

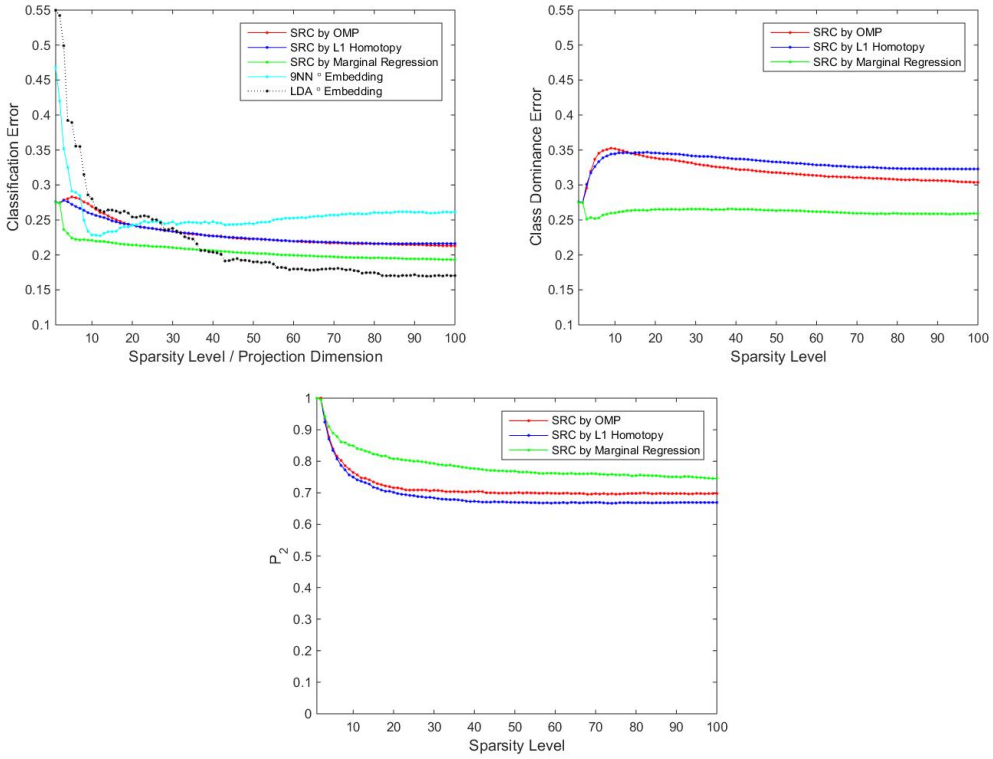

Figure 3: SRC for Wikipedia English Documents Text Feature

allows faster subset regression methods and easier data model verification for this method. Our results are illustrated by various real data analysis, including face images that roughly satisfy the subspace assumption, as well as text and network data that do not satisfy this assumption.

6. Proofs

6.1. Theorem 1 and Corollary 1

Proof. Assume that class y dominates β, we have $\|\mathcal{X}\beta_{-y}\|_2 < \|\mathcal{X}\beta_y\|_2$; furthermore, if positive class dominance holds, we have $\|\mathcal{X}\beta_{-y}\|_2 < \|\mathcal{X}\beta_y\|_2 \leq \|\mathcal{X}\beta_{-k}\|_2$ for all $k \neq y$.

Note that we can always express the testing observation as

$$\begin{aligned} x &= \mathcal{X}\beta + \epsilon \\ &= \mathcal{X}\beta_k + \mathcal{X}\beta_{-k} + \epsilon, \end{aligned}$$

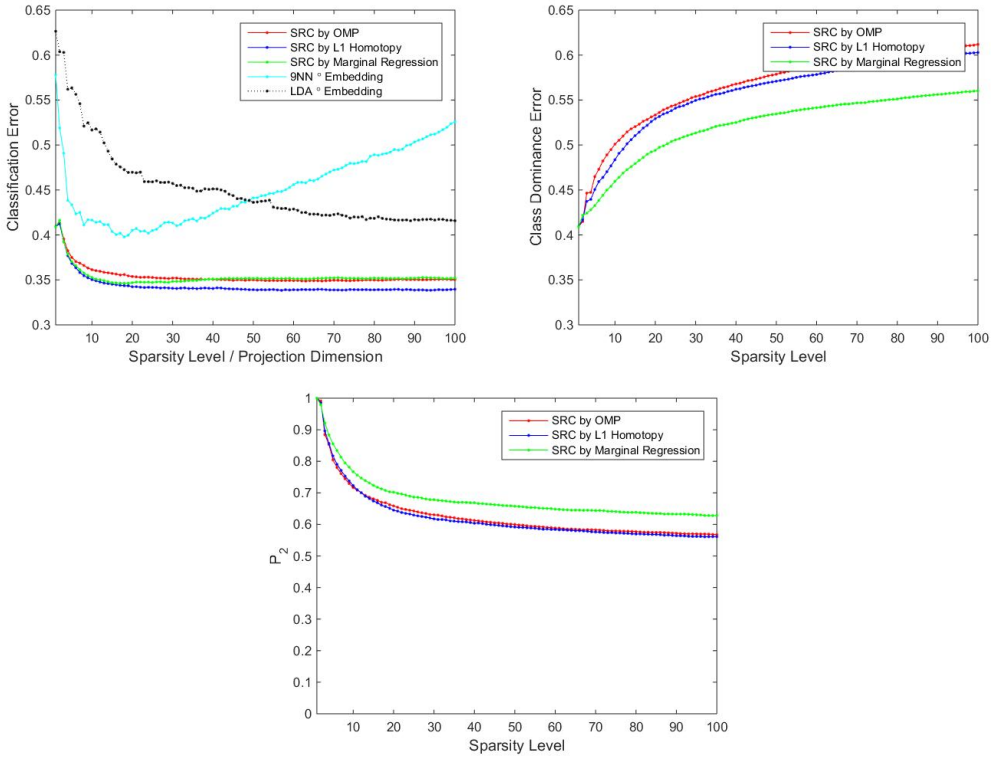

Figure 4: SRC for Wikipedia English Documents Network Feature

where ϵ is the regression residual orthogonal to both $\mathcal{X}\beta_k$ and $\mathcal{X}\beta_{-k}$ for each k, and ϵ is always fixed throughout $k = 1, \ldots, K$ for given β, $\mathcal{X}$ and x.

Thus given class dominance and positive class dominance, we have $\|\mathcal{X}\beta_{-y}\|_2 = \|x - \mathcal{X}\beta_y - \epsilon\|_2 < \|x - \mathcal{X}\beta_k - \epsilon\|_2 = \|\mathcal{X}\beta_{-k}\|_2$ for all $k \neq y$. Because ϵ is orthogonal to $x - \mathcal{X}\beta_k - \epsilon$, by the Pythagorean theorem we immediately have $\|x - \mathcal{X}\beta_y\|_2 < \|x - \mathcal{X}\beta_k\|_2$ for all $k \neq y$.

Therefore, $y = \arg\min_{k=1,\ldots,K} \|x - \mathcal{X}\beta_k\|_2$, and $g(x) = y$ for the SRC classifier.

Clearly if positive class dominance does not hold, there exist counterexamples that SRC fails to find the correct class. However, if there are only two classes (i.e., $K = 2$) or $\mathcal{X}$ and β_k are always non-negative (i.e., all observations are non-negative and the regression coefficients are constrained to be non-negative), then class y dominance guarantees that $\|\mathcal{X}\beta_y\|_2$ cannot be larger than $\|\mathcal{X}\beta_{-k}\|_2$

for all $k \neq y$. This is because

$$\begin{aligned}\|\mathcal{X}\beta_{-k}\|_2 &= \|x - \mathcal{X}\beta_k - \epsilon\|_2 \\ &= \|\mathcal{X}\beta_1 + \cdots + \mathcal{X}\beta_{k-1} + \mathcal{X}\beta_{k+1} + \cdots + \mathcal{X}\beta_K\|_2 \\ &\geq \|\mathcal{X}\beta_y\|_2,\end{aligned}$$

where the last inequality easily follows when $K = 2$ or $\mathcal{X}$ and β_k are always non-negative. Therefore, in this case class dominance implies positive class dominance, and is sufficient for correct classification of SRC. □

6.2. Theorem 2

Proof. We first decompose the testing observation as $x = \mathcal{X}_y\beta_y + \mathcal{X}_{-y}\beta_{-y} + \epsilon$, which is essentially the same as in the previous proof with a different notation for easier presentation. Note that the regression residual ϵ is orthogonal to each column of $\mathcal{X}_s$.

Next we consider the principal angle $\theta(x, \mathcal{X}_y\beta_y)$. By assuming all involved entities are positive, we have

$$\begin{aligned}\cos\theta(x, \mathcal{X}_y\beta_y) &= |x'\mathcal{X}_y\beta_y|/(\|x\|_2\|\mathcal{X}_y\beta_y\|_2) \\ &= |(\|\mathcal{X}_y\beta_y\|_2^2 + (\mathcal{X}_{-y}\beta_{-y})'\mathcal{X}_y\beta_y)|/\|\mathcal{X}_y\beta_y\|_2 \\ &= \|\mathcal{X}_y\beta_y\|_2 + (\mathcal{X}_{-y}\beta_{-y})'\mathcal{X}_y\beta_y/\|\mathcal{X}_y\beta_y\|_2 \\ &= \|\mathcal{X}_y\beta_y\|_2 + \|\mathcal{X}_{-y}\beta_{-y}\|_2 \cdot \cos\theta(\mathcal{X}_y\beta_y, \mathcal{X}_{-y}\beta_{-y}),\end{aligned}$$

where the first equality holds because $\mathcal{X}_y\beta_y$ is a vector, the second equality follows by decomposing x, and the third and fourth equalities hold when there are no negative terms involved.

Similarly, we have $\cos\theta(x, \mathcal{X}_{-y}\beta_{-y}) = \|\mathcal{X}_{-y}\beta_{-y}\|_2 + \|\mathcal{X}_y\beta_y\|_2 \cos\theta(\mathcal{X}_y\beta_y, \mathcal{X}_{-y}\beta_{-y})$.

Because $\cos\theta(\mathcal{X}_y\beta_y, \mathcal{X}_{-y}\beta_{-y})$ is always smaller than 1 (if it is 1, $\mathcal{X}_y\beta_y$ is a vector in the same direction as $\mathcal{X}_{-y}\beta_{-y}$, in which case $\mathcal{X}_s$ cannot be full rank), it is trivial to observe that $\cos\theta(x, \mathcal{X}_y\beta_y) > \cos\theta(x, \mathcal{X}_{-y}\beta_{-y})$ if and only if $\|\mathcal{X}_y\beta_y\|_2 > \|\mathcal{X}_{-y}\beta_{-y}\|_2$.

When the involved entities are not always positive, the only other possible scenario is that one absolute term negates the positive sign, e.g., $\cos\theta(x, \mathcal{X}_{-y}\beta_{-y}) = -\|\mathcal{X}_{-y}\beta_{-y}\|_2 - (\mathcal{X}_{-y}\beta_{-y})'\mathcal{X}_y\beta_y/\|\mathcal{X}_{-y}\beta_{-y}\|_2$. This can only happen when $\|\mathcal{X}_y\beta_y\|_2 > \|\mathcal{X}_{-y}\beta_{-y}\|_2$, in which case we also have $\cos\theta(x, \mathcal{X}_y\beta_y) >$

$\cos\theta(x, \mathcal{X}_{-y}\beta_{-y})$.

Therefore, class y dominates the regression vector β if and only if $\theta(x, \mathcal{X}_y\beta_y) < \theta(x, \mathcal{X}_{-y}\beta_{-y})$, assuming $\mathcal{X}_s$ is full rank. □

6.3. Theorem 3

Proof. It suffices to prove that when (x, y) satisfies the principal angle condition, the class dominance probability satisfies $P_D \to 1$.

We proceed by first assuming that $\mathcal{X}_y$ is non-empty when using $\ell 1$ homotopy. Note that $\mathcal{X}_s$ is always full rank when it is selected by $\ell 1$ homotopy.

As $\theta(x, X_1 \leq c_{xy})|Y_1 = y$ almost surely for some $c_{xy} \in [0, \pi/2)$, we always have $\theta(x, \mathcal{X}_y\beta_y) \leq c_{xy}$. And as $\theta(x, [X_1, \cdots, X_s] > c_{xy})|Y_i \neq y$ almost surely, we have $\theta(x, \mathcal{X}_{-y}\beta_{-y}) > c_{xy}$.

Therefore, with probability 1 we have $\theta(x, \mathcal{X}_y\beta_y) < \theta(x, \mathcal{X}_{-y}\beta_{-y})$, as long as $\mathcal{X}_y$ is non-empty. So it remains only to justify that $\mathcal{X}_y$ is non-empty asymptotically.

We claim that under the principal angle condition, $\mathcal{X}_y$ is asymptotically non-empty when using $\ell 1$ homotopy. First, as the prior probability of class y cannot be zero, the training data contains data of class y with probability converging to 1 as $n \to \infty$. Next, conditioning on the event that $\mathcal{X}$ contains some data of class y, the first selected datum by $\ell 1$ homotopy must be of class y (which is most correlated with the testing observation under the principal angle condition). But the first entered element may get deleted in the homotopy solution path, and it seems possible that $\mathcal{X}_y$ is empty at some s.

Let us prove this is not possible by contradiction. Suppose that at certain step s, the homotopy path deletes an element so that $\mathcal{X}_y$ is empty. Because the first added element makes $\mathcal{X}_y$ non-empty, to make sure $\mathcal{X}_y$ is empty from certain step s onwards, the deleted element $x_i \in \mathcal{X}_s$ must be the only datum of class y, i.e., $x = [x_i, \mathcal{X}_{-y}][\beta_y, \beta_{-y}]' + \epsilon$.

However, because the principal angle condition guarantees that $\theta(x, \mathcal{X}_{-y}) > c_{xy}$ and $\theta(x, x_i) \leq c_{xy}$, deleting x_i increases both $\|\epsilon\|_2$ and $\|\beta\|_1$, and can never minimize $\min_\beta \|x - \mathcal{X}\beta\|_2/2 + \lambda\|\beta\|_1$ for any $\lambda \geq 0$ (which is the objective function on the homotopy path). Thus if there is only one observation of class y remaining in the active set $\mathcal{X}_s$, that datum can never be deleted in the homotopy solution path. Thus $\mathcal{X}_y$ is almost surely asymptotically non-empty for $s \geq 1$ under the principal angle condition.

Therefore, given the principal angle condition, with probability converging to 1 we have $\theta(x, \mathcal{X}_y\beta_y) <$

$\theta(x, \mathcal{X}_{-y}\beta_{-y})$. Then if the principal angle condition holds with probability q under F_{XY}, P_D is asymptotically no less than q for $\ell 1$ minimization as $n \to \infty$. □

6.4. Corollary 2

Proof. Given the principal angle condition, class dominance holds with probability 1 asymptotically. So if the condition in Corollary 1 also holds, i.e., class dominance implies positive class dominance so that class dominance alone is sufficient for correct classification, we have $g(X) = Y$ with probability 1 asymptotically.

Therefore if those two conditions hold with probability q, the SRC error satisfies

$$\begin{aligned} L &= P_D \times P_1 + (1 - P_D) \times P_2 \\ &\to q \times 0 + (1 - q) \times P_2 \\ &\leq 1 - q. \end{aligned}$$

Furthermore, if $q \to 1$, SRC is asymptotically optimal with $L \to 0$. □

6.5. Corollary 3

Proof. Next we consider replacing $\ell 1$ minimization by other subset regression methods.

When $\ell 1$ homotopy is replaced by OMP, the only difference in our proof of Theorem 3 concerns whether $\mathcal{X}_y$ is still non-empty when using OMP. At $s = 1$, OMP adds the same element into $\mathcal{X}_y$ as $\ell 1$ homotopy, and the principal angle condition guarantees the first entered element is of class y almost surely. Unlike $\ell 1$ homotopy, OMP never deletes any element on its solution path; thus all other proofs of Theorem 3 and Corollary 2 remain the same, and OMP can achieve SRC optimality.

When $\ell 1$ homotopy is replaced by marginal regression, the first element to enter $\mathcal{X}_s$ coincides with $\ell 1$ homotopy and OMP. Therefore the principal angle condition still guarantees that $\mathcal{X}_y$ is almost surely non-empty for given $s \geq 1$. However, as marginal regression only picks s training observations that are most correlated with the testing observation, it is possible that $\mathcal{X}_s$ is no longer full rank after certain s. Thus for the proof of Theorem 3 to work, we need to constrain s so that $\mathcal{X}_s$ is full rank.

Finally, for full regression, i.e., we use $\mathcal{X}$ directly to derive the regression vector β by $\ell 2$ minimization, $\mathcal{X}_y$ is almost surely asymptotically non-empty as the prior probability of class y should

be nonzero. Therefore all proofs of Theorem 3 and Corollary 2 remain the same, and full regression can also achieve SRC optimality, as long as $\mathcal{X}$ itself is full rank. □

6.6. Corollary 4

Proof. As $X_1|(Y_1 = y) = X_{+1}I_{\theta(x,X_1)\leq c_{xy}} + X_{-1}I_{\theta(x,X_1)>c_{xy}}$, we may treat X_{-1} as from an additional class $K+1$, and keep X_{+1} still from class y.

Then the extended principal angle condition leads to the same class dominance result of Theorem 3, and Corollary 3 and Corollary 4 easily follow with essentially the same proofs. □

Acknowledgment

This work was partially supported by Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE) and the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303. The authors are also supported by the Acheson J. Duncan Fund for the Advancement of Research in Statistics, which allows us to present preliminary results of the paper at Joint Statistical Meeting, Boston, August 2014.

References

[1] J. Wright, A. Y. Yang, A. Ganesh, S. Shankar, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[2] A. Bruckstein, D. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM REVIEW*, vol. 51, no. 1, pp. 34–81, 2009.

[3] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[4] J. Yin, Z. Liu, Z. Jin, and W. Yang, "Kernel sparse representation based classification," *Neurocomputing*, vol. 77, no. 1, pp. 120–128, 2012.

[5] A. Yang, Z. Zhou, A. Ganesh, S. Sastry, and Y. Ma, "Fast l1-minimization algorithms for robust face recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3234–3246, 2013.

[6] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[8] D. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, pp. 2845–2862, 2001.

[9] D. Donoho and M. Elad, "Optimal sparse representation in general (nonorthogonal) dictionaries via l1 minimization," *Proceedings of National Academy of Science*, pp. 2197–2202, 2003.

[10] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[11] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[12] D. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm near solution approximates the sparest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 10, pp. 907–934, 2006.

[13] E. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[14] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1233, 2006.

[15] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[16] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, vol. 37, no. 1, pp. 246–270, 2009.

[17] M. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using 1 - constrained quadratic programming (lasso)," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[18] P. Belhumeur, J. Hespanda, and D. Kriegman, "Eigenfaces versus fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[19] R. Basri and D. Jacobs, "Lambertian reflection and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, pp. 218–233, 2003.

[20] R. Rigamonti, M. Brown, and V. Lepetit, "Are sparse representations really relevant for image classification?," in *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[21] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: which helps face recognition?," in *International Conference on Computer Vision (ICCV)*, 2011.

[22] Q. Shi, A. Eriksson, A. Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?," in *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[23] Y. Chi and F. Porikli, "Classification and boosting with multiple collaborative representations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1519–1531, 2013.

[24] L. Chen, C. Shen, J. Vogelstein, and C. E. Priebe, "Robust vertex classification," *submitted*, http://arxiv.org/abs/1311.5954.

[25] G. Davis, S. Mallat, and M. Avellaneda, "Greedy adaptive approximation," *Constructive Approximation*, vol. 13, pp. 57–98, 1997.

[26] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[27] L. Wasserman and K. Roeder, "High dimensional variable selection," *Annals of statistics*, vol. 37, no. 5A, pp. 2178–2201, 2009.

[28] C. Genovese, J. Lin, L. Wasserman, and Z. Yao, "A comparison of the lasso and marginal regression," *Journal of Machine Learning Research*, vol. 13, pp. 2107–2143, 2012.

[29] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[30] D. Donoho and Y. Tsaig, "Fast solution of l1-norm minimization problems when the solution may be sparse," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4789–4812, 2008.

[31] M. Osborne, B. Presnell, and B. Turlach, "A new approach to variable selection in least squares problems," *IMA Journal of Numerical Analysis*, vol. 20, pp. 389–404, 2000.

[32] M. Osborne, B. Presnell, and B. Turlach, "On the lasso and its dual," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 319–337, 2000.

[33] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[34] T. Zhang, "On the consistency of feature selection using greedy least squares regression," *Journal of Machine Learning Research*, vol. 10, pp. 555–568, 2009.

[35] T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4680–4688, 2011.

[36] D. Needle and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations of Computational Mathematics*, vol. 9, pp. 317–334, 2009.

[37] D. Needle and R. Vershynin, "Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 310–316, 2010.

[38] D. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1094–1121, 2012.

[39] M. Kolar and H. Liu, "Marginal regression for multitask learning," *Journal of Machine Learning Research W & CP*, vol. 22, pp. 647–655, 2012.

[40] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B*, vol. 70, no. 5, pp. 849–911, 2008.

[41] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional feature selection: beyond the linear model," *Journal of Machine Learning Research*, vol. 10, pp. 2013–2038, 2009.

[42] J. Fan, Y. Feng, and R. Song, "Nonparametric independence screening in sparse ultra-high-dimensional additive models," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 544–557, 2011.

[43] K. Balasubramanian, K. Yu, and G. Lebanon, "Smooth sparse coding via marginal regression for learning sparse representations," *Journal of Machine Learning Research W & CP*, vol. 28, no. 3, pp. 289–297, 2013.

[44] A. Bruckstein, M. Elad, and M. Zibulevsky, "On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4813–4820, 2008.

[45] M. Slawski and M. Hein, "Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization," *Electronic Journal of Statistics*, vol. 7, pp. 3004–3056, 2013.

[46] N. Meinshausen, "Sign-constrained least squares estimation for high-dimensional regression," *Electronic Journal of Statistics*, vol. 7, pp. 1607–1631, 2013.

[47] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[48] Y. Eldar, P. Kuppinger, and H. Bolcskei, "Compressed sensing of block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[49] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4094–4107, 2012.

[50] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[51] A. Georghiades, P. Buelhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[52] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[53] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, 2003.

[54] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[55] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.

[56] A. Singhai, "Modern information retrieval: A brief overview," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, vol. 24, no. 4, pp. 35–43, 2001.

[57] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[58] D. Cai, X. He, Y. Hu, J. Han, and T. Huang, "Learning a spatially smooth subspace for face recognition," in *Proceedings of IEEE Conference Computer Vision and Pattern Recognition Machine Learning (CVPR'07)*, 2007.